Received: 10 May 2018 | Revised: 4 September 2018 | Accepted: 7 September 2018

DOI: 10.1002/gdj3.59

DATA PAPER

WILEY Geoscience Data Journal RMetS

Shipboard automated meteorological and oceanographic system data archive: 2005–2017

Shawn R. Smith[1] | Kristen Briggs[1] | Mark A. Bourassa[1,2] | Jocelyn Elya[1] | Christopher R. Paver[3]

[1]Center for Ocean-Atmospheric Prediction Studies, The Florida State University, Tallahassee, Florida

[2]Earth, Ocean, and Atmospheric Science Department, The Florida State University, Tallahassee, Florida

[3]NOAA National Centers for Environmental Information, Silver Spring, Maryland

Correspondence
Shawn R. Smith, Center for Ocean-Atmospheric Prediction Studies, The Florida State University, Tallahassee, FL.
Email: smith@coaps.fsu.edu

Funding information
This article was funded by a grant from the Ocean Observing and Monitoring Division of the National Oceanographic and Atmospheric Administration via the Northern Gulf of Mexico Cooperative Institute administered by the Mississippi State University.

Abstract

Since 2005, the Shipboard Automated Meteorological and Oceanographic System (SAMOS) initiative has been collecting, quality-evaluating, distributing, and archiving underway navigational, meteorological, and oceanographic observations from research vessels. Herein we describe the procedures for acquiring ship and instrumental metadata and the one-minute interval observations from 44 research vessels that have contributed to the SAMOS initiative from 2005 to 2017. The overall data processing workflow and quality control procedures are documented along with data file formats and version control procedures. The SAMOS data are disseminated to the user community via web, FTP, and Thematic Real-time Environmental Distributed Data Services from both the Marine Data Center at the Florida State University and the National Centers for Environmental Information, which serves as the long-term archive for the SAMOS initiative. They have been used to address topics ranging from air-sea interaction studies, the calibration, evaluation, and development of satellite observational products, the evaluation of numerical atmospheric and ocean models, and the development of new tools and techniques for geospatial data analysis in the informatics community. Maps provide users the geospatial coverage within the SAMOS dataset, with a focus on the Essential Climate/Ocean Variables, and recommendations are made regarding which versions of the dataset should be accessed by different user communities.

KEYWORDS

data stewardship, marine meteorology, open data access, quality control, thermosalinograph

Dataset

Identifier: https://dx.doi.org/10.7289/V5QJ7F8R
Creator: Center for Ocean-Atmospheric Prediction Studies, The Florida State University, [S. R. Smith, J. J. Rolph, K. Briggs and M. A. Bourassa]
Title: Quality-Controlled Underway Oceanographic and Meteorological Data from the Center for Ocean-Atmospheric Prediction Studies (COAPS) - Shipboard Automated Meteorological and Oceanographic System (SAMOS)Publisher: National Centers for Environmental Information, NESDIS, NOAA, U.S. Department of Commerce
Publication year: 2009, ongoing
Resource type: Dataset

The Shipboard Automated Meteorological and Oceanographic System (SAMOS) initiative has been collecting, quality-evaluating, distributing, and archiving underway meteorological and oceanographic observations since 2005. The SAMOS data center, hosted by the Marine Data Center (MDC) at the Florida State University (FSU), receives data from ~30 recruited research vessels (RV) each year.

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.
© 2018 The Authors. *Geoscience Data Journal* published by Royal Meteorological Society and John Wiley & Sons Ltd.

Observations include navigational, meteorological, and surface oceanographic observations sampled at 1-min intervals (see section 2). These observations undergo scientific quality control and are distributed to a broad user community with no restrictions or holds on the data. Although this manuscript describes the SAMOS observations available from 2005 to 2017, the SAMOS initiative is an ongoing project and new observations continue to be received, processed, and archived.

The SAMOS initiative is designed to address the data needs within the air-sea interaction, satellite and remote-sensing, numerical modelling, and geoscience informatics communities. The RVs contributing to SAMOS operate in both coastal and open-ocean environments, frequently well outside the shipping lanes used by merchant vessels. As a result, SAMOS observations include the range of values collected from the tropics to the polar oceans that are desired for validating both satellite data products (e.g., May *et al*., 2017a, 2017b, Bourassa *et al*., 2003) and numerical models (e.g., Smith *et al*., 2016a). The data are also used to define the range of conditions needed to develop new satellite retrieval algorithms (Jackson and Wick, 2014). SAMOS observations can be used to estimate turbulent fluxes using bulk formulae and are the foundational dataset used to create the along-vessel track SAMOS flux product described in Smith *et al*. (2016b). These flux products can be used to evaluate numerical model flux fields. More recently, SAMOS data have been used to test new informatics concepts (Tong *et al*., 2015) and to develop a web-based Distributed Oceanographic Match-Up Service (DOMS, https://mdc.coaps.fsu.edu/doms). The latter will support future satellite calibration/validation activities.

2 | ACQUIRING SAMOS DATA

A SAMOS consists of a computerized data acquisition system, one or more navigational devices, electronic meteorological instrumentation, and oceanographic sensors (typically deployed within the vessel's scientific sea water system). The computers and instrumentation used to provide measurements to the SAMOS initiative are purchased, deployed, maintained, and operated by the RV home institution. They are not provided by the SAMOS initiative; however, the initiative does request to receive specific parameters to address the needs of our user community. Measurements made by SAMOS vary from vessel to vessel and the most common parameters provided to the MDC are listed in Table 1 along with the International System of Units (SI) used by the SAMOS initiative. Less commonly

TABLE 1 Limits outside of which SAMOS applies a bounds (B) flag for parameters commonly contributed to SAMOS

SAMOS parameter (Abbreviation)	Lower bound	Upper bound	Units
Latitude (lat)	−90	90	Degrees North
Longitude (lon)	0	359.9999	Degrees East
Speed over ground (PL_SPD)	0	15	ms^{-1}
Course over ground (PL_CRS)	0	360	Degrees
Heading (PL_HD)	0	360	Degrees
True wind direction (DIR), platform-relative wind direction (PL_WDIR)	0	360	Degrees
True wind speed (SPD), platform-relative wind speed (PL_WSPD)	0	40	ms^{-1}
Pressure (P)	950	1050	hPa
Relative humidity (RH)	0	100	percent
Air temperature (T), Wet-bulb Temperature (TW), Dewpoint Temperature (TD)	−30 (polar) −10 (mid-latitude) 10 (tropical)	15 (polar) 40 (mid-latitude) 40 (tropical)	°C °C °C
Sea temperature (TS)	−2 (polar) −2 (mid-latitude) 15 (tropical)	15 (polar) 30 (mid-latitude) 35 (tropical)	°C °C °C
Conductivity (CNDC)	0	7	Siemens m^{-1}
Salinity (SSPS)	0	50	PSU
Shortwave Radiation (downwelling; RAD_SW)	0	1400	Wm^{-2}
Longwave Radiation (downwelling; RAD_LW)	50	800	Wm^{-2}
Photosynthetically Active Atmospheric Radiation (downwelling; RAD_PAR)	0	2600	Microeinstein m^{-2} s^{-1}

This table expands on the bounds limits presented in Smith *et al*. (2016b), including more parameters quality controlled by the SAMOS initiative.

received parameters include net atmospheric radiation and measures of precipitation (volume or rate). Presently, no visual observations are provided to the SAMOS initiative—all instruments are fully automated. Every observation accepted by the MDC for inclusion in the SAMOS dataset must include the time of observation (ideally in UTC), the vessel position, and at least one additional requested parameter.

From 2005, the start of the SAMOS initiative, through 2017, the MDC recruited 44 research vessels and 30 of these vessels are still active contributors in 2018 (Table 2). The majority of these RVs are operated by the National Oceanic and Atmospheric Administration (NOAA) or universities from the U.S.; however, a partnership with the Australian Bureau of Meteorology and the Integrated Marine Observing System (IMOS) allowed recruitment of several RVs operated by Australia and New Zealand. Dates of recruitment and retirement from the SAMOS initiative are provided for each RV in Table 2.

Each RV operator is required to provide detailed vessel and instrumental metadata. The SAMOS metadata specification is based on a combination of the metadata requirements from the Voluntary Observing Ship Climate (VOSClim) program (https://www.ncdc.noaa.gov/data-access/marineocean-data/vosclim/ship-metadata; JCOMM, 2002) and the International Comprehensive Ocean-Atmosphere Data Set (ICOADS; Freeman *et al*., 2017). Key metadata requested for each vessel include the vessel name; call sign; International Maritime Organization number; operating country, name, and location of the operating institution; and contact information for the vessel operations manager and shipboard technicians. Also requested are the length, breadth, freeboard, and draught of each vessel along with digital photos of the vessel and the locations of the atmospheric and oceanographic sensors. Digital images are used by MDC personnel to identify poorly exposed sensors and to provide feedback to the operator regarding relocating sensors to improve overall data quality. For each SAMOS parameter provided to the MDC (Table 1), metadata requested for each physical device include the instrument manufacturer (make) and model number; units of observation; whether the value is measured or derived (e.g., dewpoint temperature is rarely measured but often derived from air temperature and relative humidity); whether the time stamp marks the beginning, middle, or end of the averaging period; the length of the average (in seconds); the raw data sampling rate (in Hz); the data precision; the date of instrument calibration; and the three-dimensional position of the sensor on the vessel (measured back from the bow, port/starboard of the centreline, and above/below the plimsoll line). We also request the direction convention for the winds (to or from which the wind blows), the orientation of the anemometer zero line with respect to the bow, whether the pressure value is adjusted to sea level, and the direction convention for all radiation measurements (upwelling, downwelling, or net). Additional details, definitions, and minimum requirements for SAMOS metadata are specified in Smith (2006a).

Once these metadata are received by the MDC for a newly recruited vessel, they are loaded into a ship profile MySQL database whereby the metadata can be date tracked for changes (e.g., instrument swaps, new sensors, changes in sensor location). The operators update their metadata periodically (ideally whenever instruments are changed/relocated) by either resubmitting forms to the MDC or via a graphical metadata user interface on the SAMOS website (https://samos.coaps.fsu.edu).

3 | PROCESSING WORKFLOW

The flow of SAMOS observations from the vessel to the MDC (Figure 1) begins with the operator sending all 1-min data records from the previous day to the MDC at 0000 UTC via an e-mail protocol (note: the vessels contributing to SAMOS from New Zealand and Australia post their data to a Thematic Real-time Environmental Distributed Data Services [THREDDS] server at the Australian Bureau of Meteorology and the MDC pulls the data from their server). SAMOS uses a custom key–value-paired, comma-separated value (CSV) format for data transmission (Smith, 2006b). Each operator encodes their one-minute average observations (the averaging length requested by the SAMOS initiative), derived from higher frequency (several per minute up to 1 Hz) instrumental observations, into the SAMOS format using their vessel's data acquisition software. For example, NOAA's Scientific Computer System (SCS) software has both data reduction and mailing applications that support the creation and submission of SAMOS observations to the MDC (Stepka *et al*., 2016). Once received by the MDC, these observations are converted into a network Common Data Form (netCDF; Rolph and Smith, 2005) that includes ship and instrumental metadata provided to the MDC by each operator (see section 2). The authors note that data received from a vessel in the CSV files may contain values for a part of or a day, depending on the availability of satellite transmission bandwidth or the operations plan for the vessel (e.g., the vessel may make a brief port stop during a day and send data before/after the stop in two separate files).

The data received in the CSV files then undergo a series of scientific quality control (QC) processes. The first QC procedures (section 4.1) are fully automated and results in what the MDC calls a preliminary (version 100) data file. The preliminary files, of which there may be more than one per ship and day, are available to users within 2-3 minutes of receipt of the original CSV file from each vessel.

On a 10-day delay from the observation date, intermediate (version 200) files are automatically created and made available by merging all preliminary files received for a given ship and observation day (section 4.2). This delay allows for receipt of delayed or corrected files from the RV. Finally, a select set of ships (presently all recruited NOAA vessels and the *RV Falkor*) undergo visual QC (section 4.3) to create research-quality (version 300) data

TABLE 2 Research vessels recruited to the SAMOS initiative from 2005 to 2017

Research vessel name	Call sign	Data provider	Start date	End date	No. days	No. records
Atlantis	KAQP	WHOI	01 June 2005		3,574	4,968,759
Atlantic Explorer	WDC9417	BIOS	23 July 2010		976	1,098,465
Aurora Australis	VNAA	IMOS	27 January 2008	15 March 2013	958	1,322,508
Bell M. Shimada	WTED	NOAA	24 February 2012		1,034	1,345,176
David Starr Jordan	WTDK	NOAA	19 March 2008	01 April 2009	165	207,518
Delaware II	KNBD	NOAA	03 June 2009	21 October 2010	258	336,004
Fairweather	WTEB	NOAA	01 June 2008		844	1,106,851
Falkor	ZCYL5	SOI	30 August 2013		702	915,121
Ferdinand Hassler	WTEK	NOAA	13 July 2012		339	436,848
Gordon Gunter	WTEO	NOAA	06 June 2007		1,653	2,176,841
Healy	NEPP	USCG	14 June 2007		1,001	1,350,261
Henry Bigelow	WTDF	NOAA	15 April 2007		1,498	1,874,187
Hi'Ialakai	WTEY	NOAA	01 May 2007		1,598	2,134,998
Investigator	VLMJ	IMOS	24 March 2016		402	496,339
Ka'Imimoana	WTEU	NOAA	01 June 2007	27 June 2012	999	1,331,337
Kilo Moana	WDA7827	UH	18 October 2010		1,343	1,787,195
Knorr	KCEJ	WHOI	09 May 2005	18 December 2014	2,849	3,969,565
Lawrence M. Gould	WCX7445	OPP	10 April 2007		3,229	4,630,075
McArthur II	WTEJ	NOAA	20 April 2010	05 October 2011	245	326,827
Melville	WECB	SIO	16 June 2011	08 March 2016	1,226	1,677,606
Miller Freeman	WTDM	NOAA	15 January 2007	13 October 2010	678	893,477
Nancy Foster	WTER	NOAA	26 February 2007		1,588	2,020,475
Nathaniel Palmer	WBP3210	OPP	07 November 2006		2,721	3,865,508
Neil Armstrong	WARL	WHOI	25 May 2016		461	630,988
New Horizon	WKWB	SIO	24 May 2012	29 April 2015	924	1,202,786
Oceanus	WXAQ	WHOI	12 April 2008	30 November 2011	1,059	1,443,928
Okeanos Explorer	WTDH	NOAA	17 June 2009		1,133	1,468,689
Oregon II	WTDO	NOAA	11 June 2008		1,590	2,061,329
Oscar Dyson	WTEP	NOAA	25 January 2007		2,133	2,778,740
Oscar Elton Sette	WTEE	NOAA	24 April 2009		1,363	1,816,196
Pelican	WDD6114	LUMCON	13 February 2015		127	136,411
Picses	WTDL	NOAA	13 November 2009		1,295	1,705,730
Polar Sea	NRUO	USCG	29 September 2009	13 April 2010	104	139,018
Rainier	WTEF	NOAA	03 May 2008		740	988,306
Reuben Lasker	WTEG	NOAA	09 July 2015		510	674,665
Robert G. Sproul	WSQ2674	SIO	16 April 2012		1,434	1,816,153
Roger Revelle	KAOU	SIO	13 June 2011		2,016	2,762,953
Ronald H. Brown	WTEC	NOAA	20 March 2007		1,682	2,233,158

(Continues)

TABLE 2 (Continued)

Research vessel name	Call sign	Data provider	Start date	End date	No. days	No. records
Sally Ride	WSAF	SIO	09 August 2017		133	166,120
Sikuliaq	WDG7520	UA	07 April 2015		855	1,227,684
Southern Surveyor	VLHJ	IMOS	16 April 2008	16 October 2013	959	1,228,054
Tangaroa	ZMFR	IMOS	27 April 2011		1,516	2,176,990
Thomas Jefferson	WTEA	NOAA	01 August 2012		594	764,195
Thomas G. Thompson	KTDQ	UW	29 May 2012		732	978,307

Dates indicate period when each RV contributed observations to the MDC (end date indicates retired/inactive status).
The number of days includes any day that at least a single one-minute data record was received from each vessel. The number of days and one-minute data records are totalled from the highest version daily netCDF files available for each ship. SAMOS thanks the following institutions for providing data from the vessels listed above BIOS – Bermuda Institute of Ocean Sciences, LUMCON – Louisiana Universities Marine Consortium, IMOS – Australian Integrated Marine Observing System, NOAA – National Oceanic and Atmospheric Administration: Office of Marine and Aviation Operations, OPP – Contractors for National Science Foundation Office of Polar Programs, SIO – Scripps Institution of Oceanography, SOI – Schmidt Ocean Institute, USCG – United States Coast Guard, UA – University of Alaska, UH – University of Hawaii, UW – University of Washington, and WHOI – Woods Hole Oceanographic Institution.

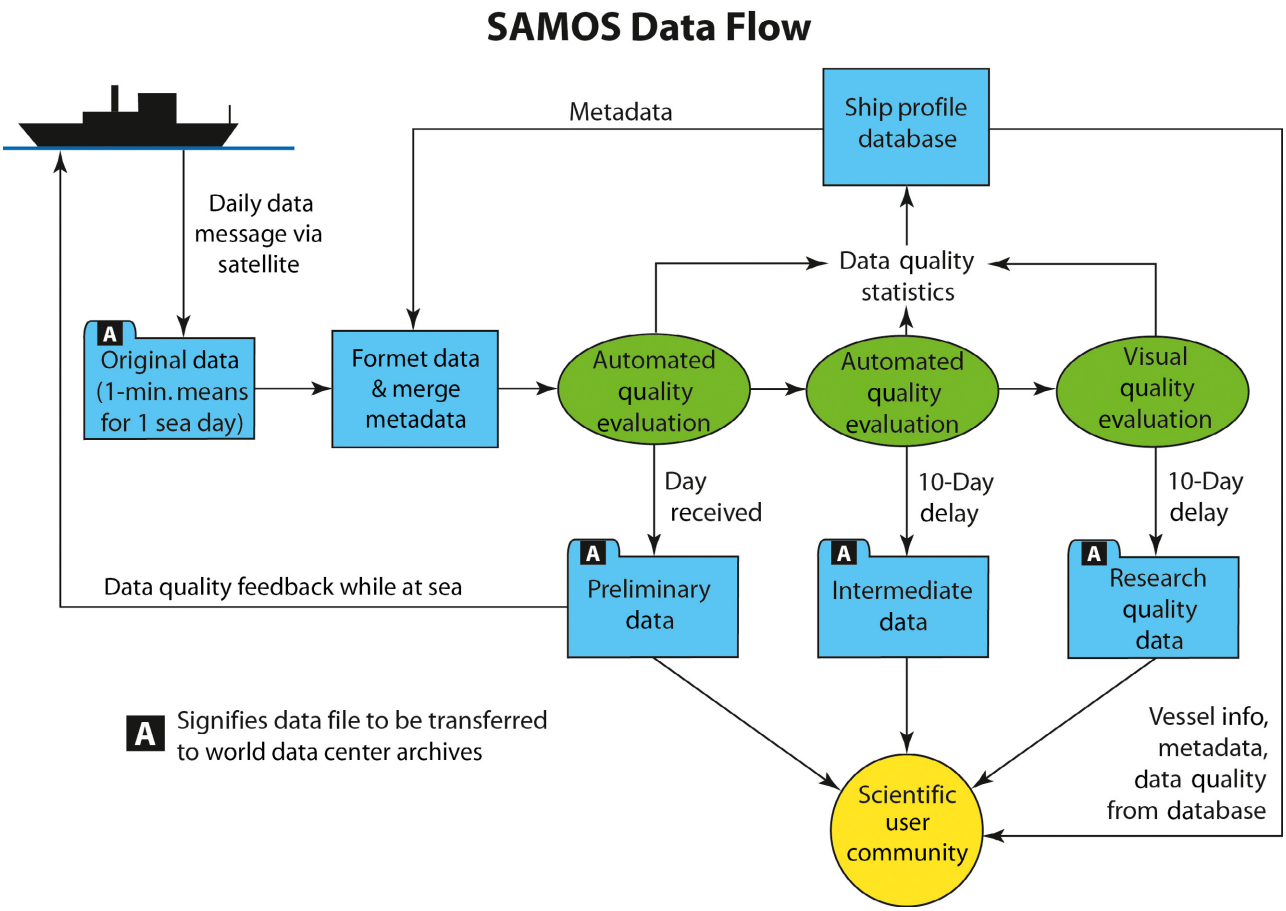

FIGURE 1 Flow of one-minute sampling rate SAMOS observations from the vessel, through the MDC, and on to the archives and user community. Each operator provides ship and instrumental metadata that is stored in the ship profile database at the MDC. These metadata are integrated into the SAMOS netCDF files and distributed to the user community via the web (Reproduced from Smith *et al*., 2016b)

files, which are available 1-3 days after the creation of the intermediate (version 200) files depending on the workload of the visual data-quality analyst.

Preliminary, intermediate, and research-quality netCDF data flles are made accessible by the MDC to users via web, FTP, and THREDDS (Unidata, 2017) services (see section 6). Each month, the original data received from the vessel and all three levels of SAMOS quality-processed data are packaged for each ship and submitted to the National Centers for Environmental Information (NCEI)—

TABLE 3 Definitions of the alphabetic flags used in the SAMOS quality control procedures

Flag	Test order	Definition
B	4[a]	Original data were out of a physically realistic range bounds outlined.
D	6	Data failed the T>=Tw>=Td test. In the free atmosphere, the value of the temperature is always greater than or equal to the wet-bulb temperature, which in turn is always greater than or equal to the dew point temperature.
E	5	Data failed the resultant wind recomputation check. When the data set includes the platform's heading, course over the ground, and speed over the ground along with platform-relative wind speed and direction, a programme recomputes the Earth-relative wind speed and direction. A failed test occurs when the difference between the reported and recomputed wind direction is >20 (or >2.5 m/s for wind speed).
F	1	Platform velocity unrealistic. Determined by comparing sequential latitude and longitude positions.
G	3	Data are greater than four standard deviations from the climatological means (da Silva et al. 1994). The test is only applied to pressure, temperature, sea temperature, relative humidity, and wind speed data.
H	v	Discontinuity (step) found in the data. Flags assigned to the maximum and minimum points in the discontinuity.
I	v	Interesting feature found in the data. Examples include: hurricanes passing stations, sharp seawater temperature gradients, strong convective events, etc.
J	v	Visual inspection shows the value to be erroneous/poor quality. The value should NOT be used.
K	v	Data suspect/use with caution – Applied when the value looks to have obvious errors, but no specific reason for the error can be determined. Some data may be useful, but uncertainty would be high and use is not recommended.
L	2	Vessel position over land based on reported latitude and longitude.
M	v	Known instrument malfunction.
N	v	Signifies that the data were collected while the vessel was in port. Typically these data, though realistic, are significantly different from open ocean conditions.
Q	v	Questionable – observation reported as questionable/uncertain in consultation with vessel operator (use with caution).
S	v	Spike in the data. Usually one or two sequential data values (sometimes up to 5 values) that are drastically out of the current data trend. Spikes occur for many reasons including power surges, typos, data logging problems, lightning strikes, etc.
Z	0	Data passed evaluation.

The order the automated QC flags are applied is noted, with visual QC flags noted with a "v" and the Z-flag denoted with a 0 as the default value at the start of the QC processing.

[a]Note the bounds test is run on latitude and longitude before the ship speed (F) and land (L) checks, but later in the processing for all other variables.

Maryland (Smith *et al.*, 2009). As a final note on SAMOS versions, sometimes it is necessary to modify existing version 100, 200, or 300 files (e.g., to fix processing errors, update metadata, change visual QC flags, etc.). In these cases, the respective version will be incremented by 001 (e.g., v201, v202, or v301, v302) and the updated files are also distributed and submitted to NCEI for archival.

4 | DATA QUALITY EVALUATION

SAMOS data QC begins with verifying that the original file came from a recruited vessel and is in the proper key–value format. Once verified, the data are converted to SI units (if necessary), checked for temporal sequence, and blended with ship and instrumental metadata (e.g., instrument height, units, sensor make/model) from the SAMOS database. This first netCDF version of the observations is not released to users until the data undergo automated QC to apply flags, resulting in a preliminary (version 100) file. SAMOS uses a hierarchical, parametric A-Z quality control scheme (e.g., each value can have only one flag; Table 3). Prior to any QC tests, all values are assumed to be correct and are assigned a Z-flag. As QC tests are run, Z-flags are replaced with an appropriate flag resulting from tests described in sections 4.1 and 4.3.

4.1 | Automated processing

SAMOS-automated QC was designed to identify and flag observations that fail tests that can be programmatically quantified. When two or more automated QC tests are run on a single data value, the tests are run sequentially giving the last run test precedence over the other QC tests. By design, the tests can flag some physically plausible values because the SAMOS QC was always envisioned to be a two-step process: automated QC followed by visual data inspection and QC by a trained meteorologist.

Automated quality tests using the ship's position data verify that (a) the vessel's speed between sequential positions is not greater than 15 m/s (platform speed check,

flag = F) and (b) the vessel is positioned over water (land check, flag = L). The platform speed test is calculated by differencing the ship's positions on a great-circle arc over a three-minute moving window and adding an F-flag to the latitude and longitude values when the calculated speed exceeds 15 m/s. The speed threshold is higher than any research vessel is likely to attain but is kept at 15 m/s as we anticipated SAMOS recruiting commercial or cruise vessels with faster cruising speeds. The three-minute window is required to ensure accurate differences between ship positions when the precision of the reported latitude and longitude data are less than 0.1°. Also, the platform speed test is run using the position data, as opposed to the ship's reported speed over the ground, because the position data must be included in a SAMOS data record (not all ships provide speed over the ground). The land check uses the vessel's latitude and longitude to look up the topographic elevation from a gridded global relief dataset. The test uses the two arc-minute ETOPO2 (National Geophysical Data Center [NGDC], 2006) for data processed prior to 31 May 2017, following which the topographic dataset was changed to the one arc-minute ETOPO1 (Amante and Eakins, 2009). Both the latitude and longitude values are assigned an L-flag when the topography value nearest the ship's position from the gridded dataset is greater than or equal to 0 m for ETOPO2 and greater than 0 m for ETOPO1 (the change to >0 m greatly reduced the number of L-flags assigned in regions of complex coastal geography). The land check (L) does not account for water bodies (lakes/rivers) that would be above sea level (e.g., the Great Lakes). In this case, L-flags will be applied to vessel positions that are over water, but when visual QC is completed these flags will be removed. We also note that vessels presently recruited to SAMOS operate in the open ocean, so such flagging for water bodies above sea-level would be rare. Finally, the land check (L) takes precedence over the platform speed check (F) when flags are applied.

Every observation is checked to ensure it falls within realistic physical limits (Table 1, flag = B). This "bounds" check uses latitude-dependent boundaries for air and sea temperatures with polar (−60°N to −90°N or 60°N to 90°N), mid-latitude (−30°N to −60°N or 30°N to 60°N), and tropical (−30°N to 30°N) ranges. The limits in Table 1 are designed to flag "likely" errors but do encompass some realistic values. For example, pressure can dip to 880 hPa in a hurricane, but the likelihood of a ship being at that location is extremely low.

The pressure, air and sea temperature, wind speed, and relative humidity also have flags applied when their value exceeds ±4 standard deviations from a monthly climatology (da Silva *et al.*, 1994; flag = G). The climatology test also uses a minimum standard deviation threshold in data sparse areas (e.g., the Southern Ocean) where da Silva *et al.* (1994) has unrealistically small standard deviations. The bounds test (B) takes precedence over the climatology test (G) when flags are applied.

Another automated QC test recalculates true winds according to Smith *et al.* (1999)—using the reported vessel course over ground, speed over ground, heading, and platform-relative wind direction and speed—and compares the resulting true winds to the true winds reported by the vessel. E-flags are applied to the reported true winds when the speed (direction) differs by more than 2.5 m/s (20°). If any of the five values required to calculate a true wind are not reported, this test is not run.

The final automated QC test ensures that the physical relationship of air temperature (T) being greater than or equal to the wet-bulb temperature (TW) being greater than or equal to the dewpoint temperature (TD) is not violated. This test is run in a pair-wise manner, checking the following in sequence: T ≥ TW, T ≥ TD, and TW ≥ TD. A D-flag is applied to both values in a pair when the relationship fails, since there is no practical way to automatically determine which of the two values in the pair is in error. The authors note that this temperature relationship test is not commonly applied to SAMOS data because moisture observations on research vessels are primarily made using relative humidity sensors (as opposed to psychrometers). When flags are applied by the automated QC processing, this multivariate temperature check is run after the bounds (B) and climatology (G) checks, so D-flags will take precedence over these other tests for the temperature values.

Completion of automated QC results in a version 100 preliminary file.

4.2 | Daily file merging

Merging multiple preliminary (version 100) files for a given ship and day removes temporal duplicates to create a single daily intermediate (version 200) file. Duplicates are identified as data records in the files being merged that have identical times. Within each duplicate record, determining which value for each parameter in those records to retain in the merged file is resolved through a series of tests that first determine whether the data values are exact, defined as a difference in value of less than 0.000001, or different. When they differ, the first test retains the value with the 'best' preliminary QC flag. Best-flag hierarchy for position data (latitude, longitude) is Z>F>L and for other parameters (sea temperature, humidity, etc.) is Z>G>E B>D, where Z is the flag used for data that do not fail any QC tests. If the flags on duplicate data values are identical, the second test compares the values in question to the 30-min mean centred on the duplicate time, retaining the value closest to the mean. If comparison to the mean fails to differentiate duplicate values, then no value for the time in

question is included in the intermediate file, since there is no way to determine which duplicate is correct. When this situation occurs, it is documented in a processing log as a compromise to allow automation of the file merge process.

4.3 | Visual evaluation

Visual QC checks on intermediate files for select vessels are completed by a trained meteorological data-quality analyst using the SAMOS Visual Data Assessment Tool (SVIDAT). SVIDAT is an interactive graphical data display and evaluation tool developed by the MDC using the Interactive Data Language (IDL; commercially available from Research Systems, Inc.). The tool supports "point and click" and "point, click, drag" activity to allow the analyst to open an intermediate SAMOS netCDF file, graphically display the data for any variable as a time series plot in a data editor window, overlay QC flags for each variable, edit the QC flags as needed, and save the edited file with a new file version. An example of assigning flags using the data editor window is shown in Figure 2. In the operational workflow, the analyst saves the first visually edited data as version 250 (used only internally) netCDF file. Subsequent internal versions (251, 252, etc.) can be created as necessary between QC editing sessions and when the analyst has completed all the required edits to the QC flags, the final step is to save the file as a research-quality (version 300) netCDF file.

In addition to the data editor window, a complementary map window exists that enables plotting the vessel's position on the globe, with or without a climatology underlay. The climatology from da Silva *et al.* (1994) includes wind speed, air temperature, sea surface temperature, precipitation, sea-level pressure, relative humidity, and total cloud amount over the oceans. Use of the map window can assist in making data flagging decisions, particularly when verifying the monthly climatology flags (G) applied by the automated QC. Both the data editor window and map window also allow for zooming in and out of a time slice or region of interest, respectively, to further aid in flag application.

Another data window in SVIDAT allows the analyst to view multiple days (up to 5) of data from one variable in a single time-series plot. This supports comparison of data from sequential dates or nearby locations, both of which offer further guidance to the analyst in making data flagging decisions.

Using SVIDAT, the analyst reviews all observations and has the option to modify/remove flags applied by the automated QC and/or add new flags based on the analyst's experience. In general, visual QC will typically involve the application of QC flags to identify discontinuities (H), interesting features (I), obviously erroneous values (J), suspicious/suspect values (K), known instrument malfunctions (M), occurrences of the vessel being in port (N), spikes (S), and questionable values resulting from external input (Q).

QC flags J, K, and S are most commonly applied by visual inspection, with K being the catchall for the various issues common to most vessels, such as (amongst others) steps in data due to platform speed changes or obstructed platform-relative wind directions, data from sensors affected by stack exhaust contamination, or data that appears out of range for the vessel's region of operation. J-flags are applied when data are obviously in error but either the error is not due to a malfunction or a malfunction has not yet been confirmed. Examples wherein a malfunction is not suspected include a thermosalinograph that is purposely denied a fresh water supply, a sensor that has iced over, or a temperature or relative humidity sensor that has been temporarily flooded by sea spray due to rough sea conditions. In practice, the analyst often must subjectively decide where to draw the line when applying suspect (K) and erroneous (J) flags. In other words, at what point does the analyst feel the gravity of the apparent error overwhelms either any measure of feasibility (this typically means the magnitude of the error) or else any potential benefit still to be gained from working with the data (e.g., salinity values that differ by several units on average from what is expected for a given region and environment of operation yet are still potentially useful for identifying salinity fronts). Spike (S) flags are applied when one to approximately five consecutive data values singularly outlie the general trend of the time series and where the values then return back to be consistent with surrounding observations, with no evidence in any other parameter that would support these outliers. An example of when a spike may represent a real weather event is a wind gust from a thunderstorm. In such a case, the wind speed may spike upwards at the same time there is a corresponding change in atmospheric pressure, temperature, and humidity that are caused by the thunderstorm gust event. In this case no S-flags would be applied to the data, but the interesting feature (I) flag might be applied by the analyst to mark the wind gust depending on how extreme a gust occurred.

Application of both suspect (K) and erroneous (J) flags requires human interaction with the data. The scenarios requiring K-flags are often too subtle and/or too ambiguous to be robustly automated, and in the case of choosing between K- and J-flags, computers are by definition not subjective. There exists a small set of J-flag circumstances that automation might handle, for example a variable that remains at a constant value in excess of a specified amount of time, but overall the data requiring J-flags do require 'eyes' on them.

S-flag application is likewise inherently difficult to automate. Where a spike exists in one variable the analyst will

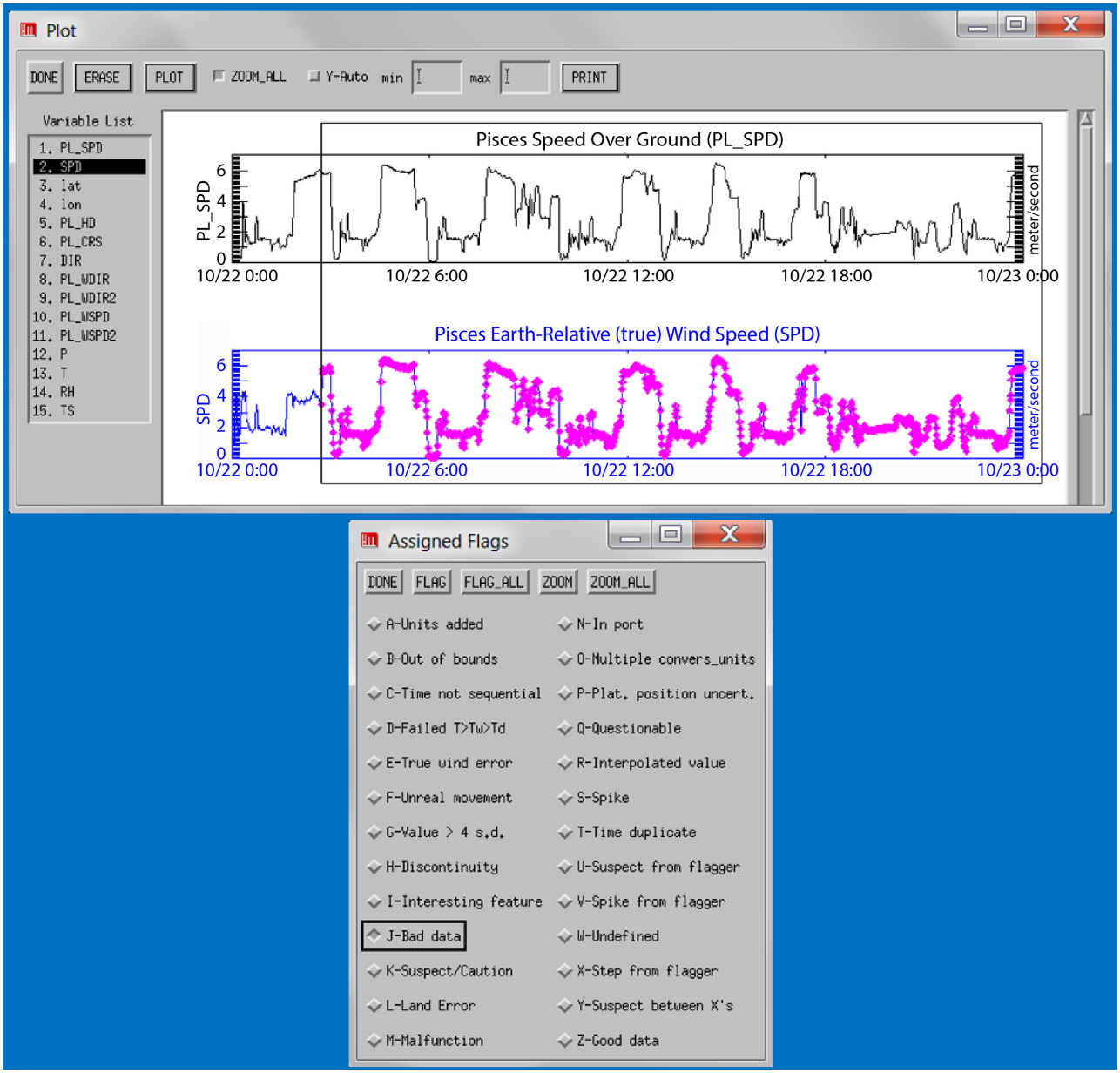

FIGURE 2 SVIDAT editor window (top), demonstrating using click-drag (rectangular outline) to highlight (pink) a portion of the Earth-relative wind speed (SPD) data (bottom plot in blue) for the *RV Pisces*. The "Assigned Flags" toggle box (bottom) pops up once a portion of data are highlighted. In this case, J-flags have been chosen and once FLAG is clicked in the upper left of the Assigned Flag window all values highlighted in pink will be assigned J-flags. When all flag assignments are complete, they are saved to the appropriately versioned netCDF file using a separate file management window. Note the editor window Variable List, which is adjusted by the user as needed, and the various options for manipulating the display atop the time series. This particular case was noted and flagged because the SPD time series was mimicing changes in the vessel speed over the ground (PL_SPD, top plot) which should not happen and was eventually identified as a platform-relative wind speed sensor malfunction (loose wiring) that resulted in incorrectly calculated Earth-relative wind speeds

sometimes also see a spike in other variables, but testing one spike against the behaviour of other variables is not always a reliable indicator. In the case of the thunderstorm wind gust noted previously, spikes in related variables are physically plausible; however, electrical interference affecting a cluster of instruments can also create coincident spikes across variables which need to be flagged. It is also difficult to determine a threshold for spike detection; they are often very large, but they also may be very small yet still visually distinct. Further, a succession of spikes can make it difficult for any automated process to differentiate between the trend and the spikes. Operating environment is another factor that complicates S-flag automation. For example, when transiting within intracoastal and inland waterways an RV often encounters a much noisier wind environment than occurs over the open ocean. Allowing

the threshold for spike detection to vary by geographic position (e.g., near-coast vs open-ocean) could be implemented, but would require significant testing and tuning of these coastal regional thresholds. In the end, "eyes" on the data are the best spike detection solution across the varying environments in which RVs operate.

On rare occasions, the analyst may also add flags marking the start and end of a discontinuity (or step) in the data. This H-flag is designed to mark the minimum and maximum points in the discontinuity, but in practice identifying and marking those individual points is very time consuming, so more often the analyst just assigns a K- or J-flag to these points, whichever is more appropriate.

Instrument malfunction (M) flags are primarily assigned when there has been communication with vessel personnel in which they have dictated or confirmed there was an actual sensor malfunction. On rare occasions, the analyst may also denote a value as questionable (Q-flag) based on communication with a vessel operator. The Q-flag is used when the concern raised about the data quality by the operator is not specifically tied to a sensor malfunction.

Port (N) flags are reserved for the latitude and longitude parameters and are infrequently applied, in an effort to minimize over-flagging of the position data. The primary application of the port flag occurs when a vessel is known to be in dry dock. The port flag may also be applied, often in conjunction with flags on other parameters, to indicate that the vessel is confirmed (visually or via operator) in port and any questionable data are likely attributable to dockside structural interference, although this practice is traditionally only used in extreme cases.

The interesting feature (I) flag is a unique characteristic of the SAMOS QC system. This flag allows the analyst to optionally identify interesting meteorological or oceanographic features in the data (e.g., pressure minima associated with a frontal passage or tropical cyclone). No set criteria exist for assigning the I-flag, just the subjective experience of the analyst to note interesting events.

SAMOS data analysts may also apply Z-flags to data, in effect removing flags that were applied by automated QC. For example, bounds testing (B-flag) is dependent on latitude for air and sea temperature (Table 1), and occasionally a realistic value is assigned a B-flag simply because the observation was made near one of the latitude boundaries used in the test. Case in point, sea temperature in the extreme northern Gulf of Mexico periodically achieves values of 32°C or 33°C in the summer, but portions of the Gulf are north of 30° latitude and thus fall into a region where such high temperatures are coded as 'out of bounds'. In this case the B-flags would be removed by the data analyst and replaced with good data (Z) flags.

5 | SAMOS OBSERVATIONS: 2005–2017

Since the inception of the SAMOS initiative in 2005, the MDC has processed over 68 million one-minute data reports provided by 44 research vessels. These data are

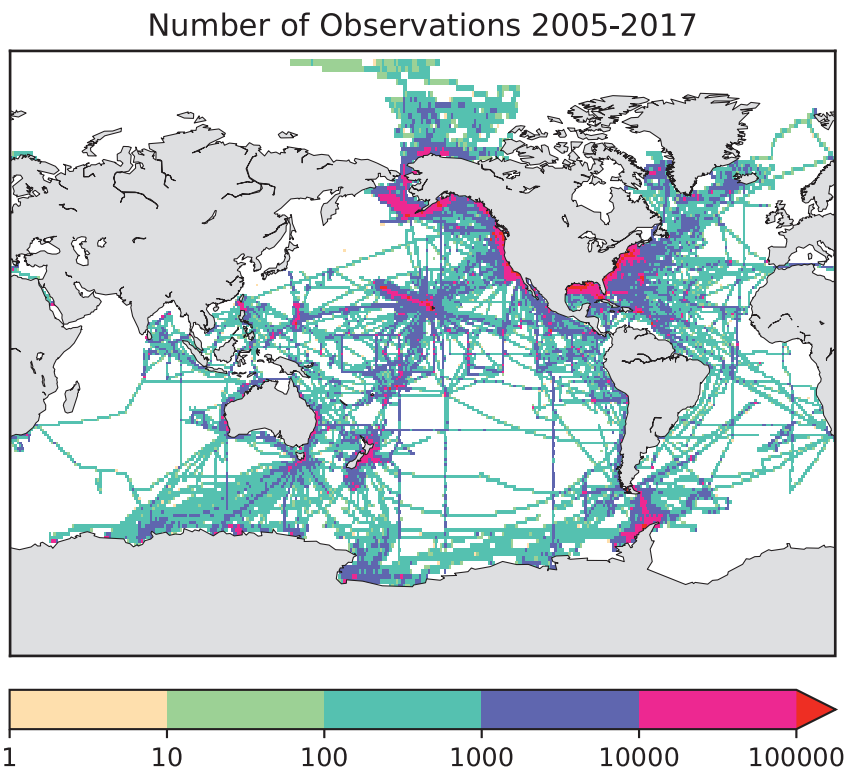

FIGURE 3 Total number of SAMOS one-minute data reports from 2005 to 2017. Values are binned in 1-degree latitude by 1-degree longitude cells with magnitude coded according to the colour bar

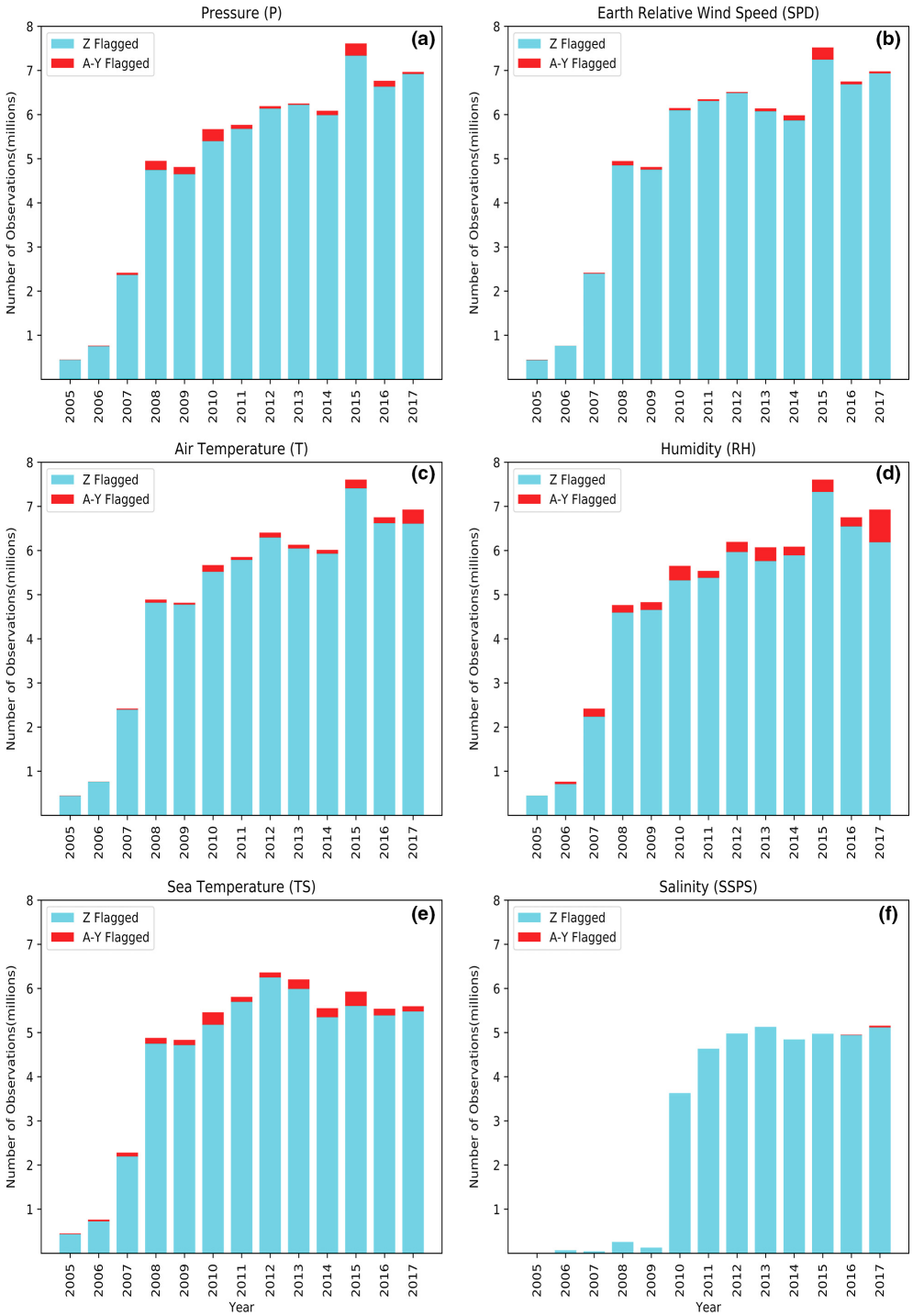

FIGURE 4 Total number of SAMOS observations from 2005 to 2017 for the following ECV/EOV (a) pressure, (b) wind speed, (c) air temperature, (d) relative humidity, (e) sea surface temperature and (f) sea surface salinity. Each bar shows the number of records that passed all (Z) or failed one or more (A-Y) of the SAMOS automated and visual QC tests. Observation counts are from the highest quality-controlled version (intermediate or research) daily netCDF files available for each ship. The rapid increase in observations from 2005 to 2008 (2010 for salinity) is a result of recruiting ships to provide each parameter early in the SAMOS initiative. The fraction of data failing one or more QC test has remained stable for each variable over the observing period

collected across the global oceans, with the highest observation density being around North America (Figure 3). The coverage varies from year to year since SAMOS relies on vessels conducting research cruises with varying missions. Although this results in some regions with sparse data coverage, the trade-off is that SAMOS observations are frequently made in data-sparse locations, which are of high value to our science users.

SAMOS data collection also focuses on parameters that have been identified by the global climate and ocean community as Essential Climate/Ocean Variables (ECV/EOV, Bojinski *et al*., 2014). Routinely collected SAMOS ECVs include pressure, wind speed and direction, air temperature, and water vapour (typically from relative humidity measurements), and EOVs include sea surface temperature and sea surface salinity. Following initial vessel recruitment from 2005 to 2007, the number of ECV/EOV observations made by the SAMOS fleet has remained stable (Figure 4). Select vessels also measure precipitation and components of the surface radiation budget (both ECVs). Finally, SAMOS observations frequently include all the parameters required to estimate ocean surface heat flux and wind stress (both EOVs) by applying bulk formulae for air-sea fluxes (e.g., Smith *et al*., 2016b).

Overall, the SAMOS QC processing applies flags to 4.9% of the observations. The QC flags applied to the SAMOS data vary by year and ship. Automated QC typically applies fewer flags (~2% of observations) than are added during visual QC (which results in ~5% more observations being flagged), noting that visual QC is only applied to a subset of the recruited RVs.

The authors note that conducting visual QC adds value to the resulting data products available to the user community. As a result of funding cuts, the MDC discontinued visual QC for several recruited vessels in 2013. The result was a significant drop in the number of quality flags being assigned to the data from these vessels. For example, the percent of all data assigned A-Y flags for *L. M. Gould* dropped from 10.2% in 2012% to 1.3% in 2013, with several common data problems not being flagged in 2013 (Figure 5). From a user perspective, the lower flag percentage after 2013 may imply the data are higher quality, when in fact a number of problems exist in the data that are not identified when only automated QC is applied. This is why we recommend using the research-quality SAMOS netCDF files whenever they are available.

6 | DATASET ACCESS AND USAGE

The MDC at FSU has a long-standing policy to ensure free and open access to underway marine meteorological observations. For SAMOS, the policy ensures that all data provided to FSU will be redistributed to the user community and national archive centers without any restrictions or 'holds'. Data providers are notified of this policy and can suspend data transmissions from their ship to the MDC when requested/required for any given cruise. For example, if a vessel is conducting classified operations or working in a nation's exclusive economic zone, they likely will not transmit SAMOS data to the MDC.

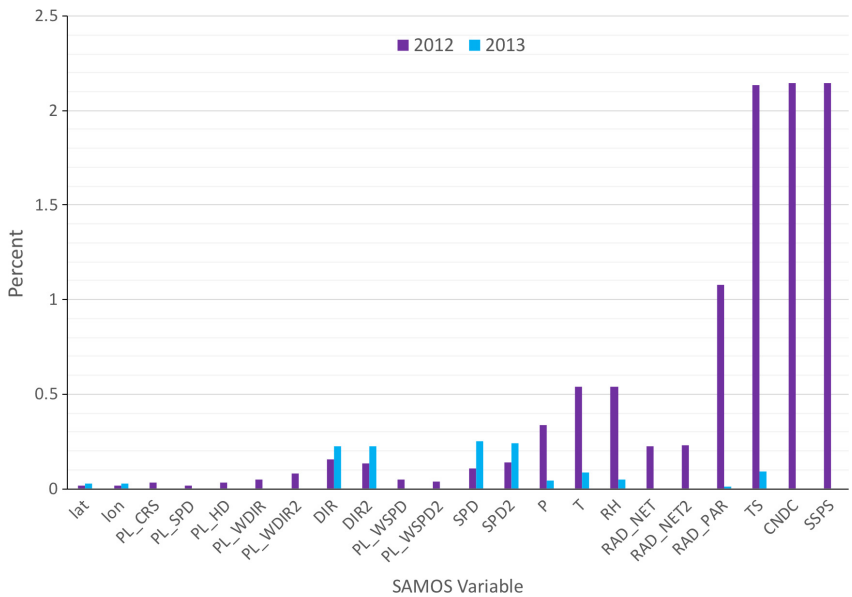

FIGURE 5 Comparison between total percentage of data flagged for each variable in 2012 and 2013 for the *Laurence M. Gould*. *L.M. Gould* personnel typically turn off the sea water intake pump when the vessel is within sea ice or in port. In 2012, when such conditions occurred a SAMOS data analyst J-flagged *L.M. Gould's* internal thermosalinograph data (i.e., pumped-water sea temperature [TS], conductivity [CNDC], and salinity [SSPS]). A case of +100 μE/m^{2}s-biased photosynthetically active radiation (RAD_PAR) was also manually K-flagged in 2012. Aside from these distinctive 2012 flag circumstances, the *L.M. Gould's* SAMOS measurements, particularly pressure (P), air temperature (T), and relative humidity (RH) were routinely impacted by complex wind flow patterns over and around the superstructure of the ship resulting in additional visually-applied flags. Post 2012, when visual QC was discontinued for the *L.M. Gould*, any events similar to those described here were overlooked by automated QC

The primary archive for the SAMOS dataset is at NCEI (https://dx.doi.org/10.7289/V5QJ7F8R). NCEI provides access via a granule search or direct data downloads using THREDDS, Open-source Project for a Network Data Access Protocol (OPeNDAP), HTTPS, or FTP.

Additional data search, access, and download services are provided by the MDC for all SAMOS observations. Data can be accessed on the web by observation date at https://samos.coaps.fsu.edu/html/data_availability.php or by cruise identifier at https://samos.coaps.fsu.edu/html/cruise_data_availability.php. Note that data access by cruise is only available for a select set of vessels whose cruises are catalogued by the Rolling Deck to Repository (R2R; https://rvdata.us) project. The MDC also maintains THREDDS (https://tds.coaps.fsu.edu/thredds/catalog_samos.html) and FTP (ftp://ftp.coaps.fsu.edu/samos_pub/data) services for users.

The preliminary, intermediate, and research-quality SAMOS data are provided to users in netCDF (Rolph and Smith, 2005). These netCDF files were developed referencing the Cooperative Ocean/Atmosphere Research Data Service (COARDS, 1995) netCDF conventions, though many of the variable names are custom to the SAMOS format. All three levels of QC'd data are available from the MDC and NCEI. The original key–value pair CSV file received from the vessel is also available from the archives at NCEI.

The MDC staff recommends use of the research-quality (version 300 or greater) dataset whenever it is available and the intermediate (version 200) dataset for vessels where the research-quality product is not created. The preliminary product is recommended primarily for operational activities that require data access soon after the data are received at the MDC (for those with requirements to use the data prior to the 10-day delay when the intermediate product is created). Users of the preliminary data are cautioned that they will encounter multiple files for the same ship and day, many with duplicate records.

OPEN PRACTICES

This article has earned an Open Data badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available at http://dx.doi.org/10.7289/V5QJ7F8R. Learn more about the Open Practices badges from the Center for Open Science: https://osf.io/tvyxz/wiki.

ACKNOWLEDGEMENTS

Preparation of this article and base support for the SAMOS project are provided by a cooperative agreement (NA16OAR4320199) from the Climate Program Office, Ocean Observing and Monitoring Division of the National Oceanographic and Atmospheric Administration (FundRef #100007298) via a subaward (191001.363513.01D) from the Northern Gulf of Mexico Cooperative Institute administered by the Mississippi State University. Additional support for SAMOS is provided by the National Science Foundation's Oceanographic and Technical Services Program (award OCE-1447797), the Office of Naval Research (via NSF), and via annual contracts from the Schmidt Ocean Institute. The authors wish to thank the marine technicians, captains, crew, and shore-side personnel at the research vessel operating institutions without whom the SAMOS project could not be successful. We also thank Jeremy Rolph, Jiraporn Whalley, James Stricherz, James Lamm, Jiangyi Hu, Kris Suchdeve, E. Michael McDonald, and all the student programmers that have worked at the MDC over the years developing and improving the SAMOS data processing software. The authors also thank Bethany Sanders who developed several of the graphics used in the manuscript.

ORCID

Shawn R. Smith http://orcid.org/0000-0003-1392-3077
Mark A. Bourassa http://orcid.org/0000-0003-3345-9531
Jocelyn Elya http://orcid.org/0000-0002-2261-1449
Christopher R. Paver http://orcid.org/0000-0002-5211-7820

REFERENCES

Amante, C. and Eakins, B. W. (2009). ETOPO1 1 Arc-Minute Global Relief Model: Procedures, Data Sources and Analysis. NOAA Technical Memorandum NESDIS NGDC-24. National Geophysical Data Center, NOAA. doi:https://doi.org/10.7289/V5C8276M [12 December 2016].

Bojinski, S., Verstraete, M., Peterson, T. C., Richter, C., Simmons, A. and Zemp, M. (2014) The concept of essential climate variables in support of climate research, applications, and policy. *Bulletin of the American Meteorological Society*, *95*, 1431–1443. https://doi.org/10.1175/BAMS-D-13-00047.1.

Bourassa, M.A., Legler, D.M., O'Brien, J. J. and Smith, S.R. (2003) SeaWinds validation with research vessels. *Journal of Geophysical Research*, *108*, https://doi.org/10.1029/2001JC001028.

COARDS (1995). COARDS NetCDF Conventions. Available at: https://ferret.pmel.noaa.gov/Ferret/documentation/coards-netcdf-conventions [Accessed 11 July 2018].

daSilva, A.M., Young, C.C., & Levitus, S. (1994) Atlas of Surface Marine Data, Vol. 1: Algorithms and Procedures. NOAA Atlas NESDIS 6, U. S. Department of Commerce, NOAA, 83pp. Available from: NOAA/NODC, Customer Service, E/OC, 1315 East-West Highway, Silver Spring, MD 20910.

Freeman, E., Woodruff, S.D., Worley, S.J., Lubker, S.J., Kent, E.C., Angel, W.E. *et al* (2017) ICOADS Release 3.0: a major update to the historical marine climate record. *International Journal of Climatology*, *37*, 2211–2232. https://doi.org/10.1002/joc.4775.

Jackson, D.L. and Wick, G.A. (2014). Propagation of uncertainty analysis of CO2 transfer velocities derived from the COARE gas transfer model using satellite inputs. *Journal of Geophysical Research: Oceans*, *119.3*, 1828–1842. https://doi.org/10.1002/2013JC009271.

Joint World Meteorological Organization (WMO)-Intergovernmental Oceanographic Commission (IOC) Technical Commission for Oceanography and Marine Meteorology (JCOMM). (2002) Voluntary observing ships (VOS) climate subset project (VOSClim) project document, revision 2. WMO/TD-No. 1122. Geneva: WMO.

May, J.C., Rowley, C. and Barron, C.N. (2017a). NFLUX satellite-based surface radiative heat fluxes. Part I: Swath-Level Products. *Journal of Applied Meteorology and Climatology*, *56*, 1025–1041.

May, J.C., Rowley, C. and Barron, C.N. (2017b). NFLUX satellite-based surface radiative heat fluxes. Part II: Gridded Products. *Journal of Applied Meteorology and Climatology*, *56*, 1043–1057.

National Geophysical Data Center. (2006) 2-minute Gridded Global Relief Data (ETOPO2) v2. National Geophysical Data Center, NOAA, https://doi.org/10.7289/V5J1012Q [26 August 2003].

Rolph, J.J. and Smith, S.R. (2005). *SAMOS netCDF code manual for quality controlled surface meteorology data. Report 05–03* (p. 35). Center for Ocean-Atmospheric Prediction Studies, Tallahassee, FL, 32306–2741. Available at: https://samos.coaps.fsu.edu/html/docs/samos_netcdf_manual.pdf [accessed 12 July 2018].

Smith, S.R., Bourassa, M.A. and Sharp, R.J. (1999). Establishing more truth in true winds. *Journal of Atmospheric and Oceanic Technology*, *16*, 939–952.

Smith, S.R., Briggs, K., Lopez, N. and Kourafalou, V. (2016a) Numerical model evaluation using automated underway ship observations. *Journal of Atmospheric and Oceanic Technology*, *33*, 409–428. https://doi.org/10.1175/JTECH-D-15-0052.1.

Smith, S.R., Rolph, J.J., Briggs, K. and Bourassa, M.A. (2009) Quality-controlled underway oceanographic and meteorological data from the Center for Ocean-Atmospheric Prediction Studies (COAPS) - Shipboard Automated Meteorological and Oceanographic System (SAMOS). National Oceanographic Data Center, NOAA, https://doi.org/10.7289/V5QJ7F8R.

Smith, S. R., Lopez, N. and Bourassa, M. A. (2016b) SAMOS air-sea fluxes: 2005–2014. *Geoscience Data Journal*, https://doi.org/10.1002/gdj3.34.

Smith, S.R. (2006a). SAMOS Metadata Form Instructions. Revision 2, 10 pp. Center for Ocean-Atmospheric Prediction Studies, Tallahassee, FL, 32306–2741. Available at: https://samos.coaps.fsu.edu/html/recruit/SAMOS_metadata_op02.pdf [accessed 12 July 2018].

Smith, S.R. (2006b) SAMOS Data Exchange Format. Revision 3, 7 pp. Center for Ocean-Atmospheric Prediction Studies, Tallahassee, FL, 32306–2741. Available at: https://samos.coaps.fsu.edu/html/recruit/SAMOS_data_exchange_op03.pdf [accessed 12 July 2018].

Stepka, T., Shields, D., Chang, S., Cromer, K., Katebini, J. and Zublay, P. (2016) Scientific Computer System Version 4.9 User's Guide. 366 pp. Available from National Oceanic and Atmospheric Administration, Office of Marine and Aviation Operations, Information Management Division, Software Development Group, 8403 Colesville Rd, Suite 500, Silver Spring, MD 20910.

Tong, Y., Zhang, X. and Chen, L. (2015) Tracking frequent items over distributed probabilistic data. *World Wide Web*, *1–26*, https://doi.org/10.1007/s11280-015-0341-5.

Unidata (2017) THREDDS Data Server [V4.6.11, build 2017–12-04]. Boulder, CO: UCAR/Unidata Program Center. https://doi.org/10.5065/D6N014KG.

How to cite this article: Smith SR, Briggs K, Bourassa MA, Elya J, Paver CR. Shipboard automated meteorological and oceanographic system data archive: 2005–2017. *Geosci Data J*. 2018;5:73–86. https://doi.org/10.1002/gdj3.59